LangEdu Journal
Vol. - No. -, Month-Year, pp.

AN ANALYSIS STUDY: DO NATIONAL EXAMINATION ITEMS OF ENGLISH SUBJECT CONTAIN TEST-WISENESS?

Gunawati Dwi Utami
English Language Teaching Study Program
Postgraduate School, Universitas Islam Malang, Indonesia
Email: faniamil@yahoo.com

Abstract

This study aimed to analyse the test-wiseness of three sets of national examination items for junior high school students, the year of 2016, 2017, and 2018. The reasearcher investigated which test-wiseness subscale(s) appeared, how frequent the test-wiseness appeared, and how item validity, item difficulty, item discrimination, distractor effectiveness and the reliability were related to the items having the test-wiseness for the students studying at higher, middle and lower schools at Malang city. The research design was quantitative evaluation research. The steps to conduct this research were 1) determining test-wiseness subscales for ESL/EFL; 2) analysing the data based on the subscales; 3) counting the appeared subscales and the frequency; 4) checking the item analysis; and 5) drawing conclusion. The findings showed that similar-option, stem-option and item give-away were the appeared subscales. There were 15,3% items having test-wiseness. The statistical analysis showed that the items having test-wiseness did not have any problems with the validity, the difficulty index, the distractors effectiveness, except from those from the lower level school. This meant that test-wiseness did exist in national examination items and the appearance affected the students, especially those coming from high level school who had good cognitive ability. It was suggested to the test developer of national examination and English teachers to carefully develop a test, considering the three test-wiseness subscales, namely stem option, similar-option, and item giveaway. The headmaster and English teachers community should hold a teacher training on assessment. For the future researcher, analysing the test-wiseness of the future national examination items with many more new sources and broader population is still worthly needed.

Keywords: test-wiseness, national examination items, English subject.

INTRODUCTION

National examination of English subject for junior high school students consists of 50 multiple choice items. The usage of multiple choice test items makes it possible to test many topic areas than the essay test items do, as stated by Seaman (2003). Another benefit of using multiple choice test items is that it can measure various cognitive level (*Kementerian*, 2017). Besides, it is marked more

easily, effectively and efficiently. Furthermore, the reliability of multiple choice tests is assured for the assessment is done objectively (Sulistyo and Rahmajanti, 2003).

Having many advantages, the usage of multiple choice tests also has some weaknesses. One of them is that there are chances to guess or to respond advantageously the correct answer although the students do not have any related knowledge about the material being tested (*Kementerian,* 2017).

This gets along with the students' complaints. Frequently, the result of national examination was dissapointing some students who have made great effort to prepare the test. They got lower score than their expectation, whereas their friends who made little effort or even did not make any effort to prepare the exam or did not seriously follow the teaching learning activities got better score. Moreover, some students also complained the different score they got eventhough they had the same ability with their friends.

The ability to respond advantageously to multiple choice items to obtain credit without knowledge of the subject matter being tested is called test-wiseness (Evans, 1984). It means that test-wiseness possibly exists in any multiple choice tests.

Moreover, Evans (1984) states that even though professional test constructors are more familiar with test-wiseness principles and are more precise in terms of item-writing skills, studies have shown that many standardized tests contain systematic biases in the keying of correct answers. National examination is a standardized-test for Indonesian students as the test items were written by national test item writer team chosen by *Badan Standar Nasional Pendidikan (BNSP)* of Indonesian Ministry of Education. It means that as standardized test for junior high school students, there is a possibility for national examination items to have test-wiseness.

In addition, Allan explained that test-wiseness can be "a source of invalidity" of tests (1992). High validity gives the accurate score, showing the true performance of test-takers, while the lower validity gives inaccurate score, showing untrue performance of test-taker. It means if the test-wiseness exists in the national examination items, it will mislead the score of English subject.

The mismatch between the students' true ability and their score became problem since national examination score is data to determine where they are going to continue thier study. Furthermore, national examination score nationally is used as one of the basis of government-made decision as stated in *Permendikbud* no 23/2016.

Considering the important role of national examination score as explained previously, the researcher investigated whether or not the national examination items were free of test-wiseness. If any, the researcher wanted to know which subscales of test-wiseness appeared and how the frequency was.

Being curious to know those problems, the researcher analysed the national examination items of academic year of 2016, 2017, and 2018 consisiting of 50 items of each, which meant there were 150 items being investigated.

Besides, to know whether or not the items containing test-wiseness influenced the students' answer, the researcher administered the test items to grade 9 students of public junior high schools at Malang city, catagorized as higher level school, middle level school and lower level schools, namely SMP Negeri X Malang, SMP Negeri Y Malang and SMP Negeri Z Malang respectively.

The researcher's observation showed that the students who were studying in higher level school have higher ability in English subject, and those who were in the middle school level have middle ability, while the lower school have lower ability. Administering the national test to the three school levels, the reseacher expected to know further whether or not the test-wiseness influenced students in different levels of schools.

The researcher formulated the research problems into three questions: 1) what test-wiseness subscales appear on national examination items of English subject?; 2) how frequent is the test-wiseness appearance on national examination items of English subject?; 3) how are item validity, item difficulty, item discrimination, distractor effectiveness and the reliability based on the test-wiseness subscales of the students studying at higher, middle and lower level schools at Malang city?

The objectives are 1) to investigate what test-wiseness subscales appear on national examination items of English subject; 2) to investigate the frequency of test-wiseness appearance on national examination items of English subject; 3) to investigate how item validity, item difficulty, item discrimination, distractor effectiveness and the reliability are based on the test-wiseness subscales of the students studying at higher, middle and lower level schools at Malang city.

METHOD

The research design was quantitative evaluational research. The concept of population and sample were applied in two ways, they were the investigated object, i.e. the test items, and the subject doing the test. The former was related to national examination items, the latter was related to the students doing the national examination items.

The national examination items being investigated in this study were the last three years (2018, 2017 and 2016) national examination items as the population. They consisted of around 20 sets of each year. Then, it was taken one set of the test items randomly each year. Thus, the sample of the investigated object was three sets of the last three years national examination items.

In relation to the subject doing the test, the national examination items were administered into grade 9 students of junior high schools at Malang city. There are 27 public junior high schools in the city whose 7.139 grade 9 students.

The researcher used clustered random sampling in this study. She classified the schools based on their average score of National Examination 2017/2018 into 3 levels of school, namely 1) higher level school, 2) middle level school, and 3) lower level school. If the average score was more than 80, the school belong to higher level school, while the middle level school gained average score between 60 to 80. The lower level was for the school which gained average score of less than 60.

Table 3.1 School Classification Based on National Examination Score

No	Classification	Schools
1	Higher Level School (>80)	SMPN A, B, X
2	Middle Level School (60 to 80)	SMPN C, D, E, F, G, H, I, J, K, L, M, N, O, Y, P, Q.
3	Lower Level School (<60)	SMPN R, S, T, Z, U, V, W.

Based on the classification, the researcher took one school of each level which was reachable for the researcher to administer. This means there were 3 schools as the sample, they were SMPN X, SMPN Y and SMPN Z respectively for each level.

From each school, the researcher took 3 classes of grade 9 which was chosen randomly to administer the test: 1) At SMPN X Malang, she got class 9-A, 9-H, and 9-I, consisting of 32 students of each, 2) SMPN Y Malang, she got classes of 9-2, 9-3 and 9-5 consisting of 34 students of each class, and 3) at SMPN Z Malang, the taken sample classes were 9-B, 9-C, and 9-D, consisting of 29 students of each. Thus, the total number of students involved in this study were 285 students of grade 9 from 3 levels of public junior high schools at Malang city.

Every school got 3 sets of national examination items of the last 3 years (2016, 2017 and 2018). Therefore, each class did 1 set of the national examination items. The test was administered at the 3 schools in March 2019.

The research instrument of this study was 150 multiple-choice items of national examination of the last three years, consisting of 50 items for each. It means there are 150 items totally.

In conducting this study, the researcher used series of steps. They were:

1) determining test-wiseness subscales for ESL/EFL;
2) analysing the data based on the test-wiseness subscale using a table;
3) counting:
   a) Test-wiseness subscale which is the most frequently appeared;
   b) the frequency of appeared test-wiseness;

4) checking the item analysis manually or digitally using iteman version 3.5 based on the test-wiseness subscales of each school level; and
5) drawing conclusion.

In collecting data of this study, the researcher chose which national examination items were going to be used, as this was the main issue of this study. Being curious whether or not the test-wiseness existed in the national examination items, the researcher took the items of year 2016, 2017 and 2018.

Next, to have the data to be analysed, the test items should be administered into the chosen sample students. Then, the researcher asked for each school English teacher's help to administer each set of test items to each sample class.

As there were many theories of test-wiseness principles, firstly the researcher selected the test-wiseness subscales which were going to be used in this study. As explained previously, the researcher took four test-wiseness subscales for ESL/EFL formulated by Allan, they were stem option, grammatical cue, similar option, and item giveaway.

Secondly, she started to analyse the last three years national examination items by inputing 150 items into the provided table. Each item was analysed by putting check (√) for item containing one or more test-wiseness, whereas the researcher put dash (-) when there is no test-wiseness in it. It should be done for all items, totally 150 items.

In order to know test-wiseness subscale which was the most frequently appeared in the national examination items, the reseacher counted the number of test-wiseness for each subscale, it was devided by the total item, then the result was multiplied 100%. Based on the percentage of each subscale, the researcher got the data of test-wiseness subscale which were appeared in the national examination items..

After that, the researcher counts the frequency. In order to know the frequency of the test-wiseness appeared in the national examination items of English subject in the years of 2016, 2017, 2018, the researcher used the main data analysis (step 2). The data were presented in form of percentage, derived from the existing test-wiseness, devided by the total items, then multiplied 100%. In short the formula is follows:

$$\frac{\textbf{Amount of Test-wiseness}}{\textbf{Amount of total items}} \textbf{ x 100\%}$$

To understand the data of frequency easily, the researcher used a table consisting of year, number of items, number of items having test-wiseness, and percentage. Then, the total of each was counted.

The next step was checking the item analysis based each test-wiseness subscale of the school level. The researcher investigated the index of each test-wiseness subscale of higher, middle and lower school. They were compared whether or not the students coming from higher level school had higher achievement than those who were studying at middle level school, and those from middle school were better achieved than those from lower level school.

In the purpose of knowing the item validity, the item difficulty, item discrimination, distractor effectiveness, and item reliability of the items having the test-wiseness, the researcher counted manually or digitally using ITEMAN (Item and Test Analysis) version 3.50.

To know validity of the items having test-wiseness, the reasearcher used catagory by Arikunto (2016), as in table 3.2. According to the Manual of ITEMAN (2006), the point biserial in Iteman is the validity value of the test item. Using the point biserial in Iteman, the researcher investigated the validity value of items having test-wiseness.

Table 3.2 Catagories of Item Validity

Catagories	Validity Value
Very Low	0.00 – 0.20
Low	0.21 – 0.40
Sufficient	0.41 – 0.60
High	0.61 – 0.80
Very High	0.80 – 1.00

(Arikunto 2016:89)

To know the item difficulty, the difficulty index was classified into three catagories, as stated in Table 3.3. The researcher found out the difficulty index of the items containing test-wiseness by counting the proportion of the students' correct answer of each items towards the total number of the test-takers.

Manually, the researcher checked the difficulty level using percentage of the students' correct answer of each subscale of the three school level. Through this data, the researcher knew the link between the test items having the test-wiseness and the students' answer of each item. The data were presented in form of percentage for each items of each school for each subscale. The more the percentage of the correct answer, the items were easier for the students in a certain school level. Then, the researcher compared the average of percentage of each subscale among the three levels of the students. The catagory of the difficulty level was shown in Table 3.3

Table 3.3 Catagories of Difficulty Index

Catagories	Difficulty Index
Difficult	0.00 – 0.30
Average	0.31 – 0.70
Easy	0.71 – 1.00

(Arikunto,2016:225)

Next, the reseacher analysed the item discrimination index. Discrimination has a function to distinguish between the students who master the competence being tested and those who do not master ones (Ariyana, 2011). In iteman, classical discrimination index is shown by the score of Mean-Item Tot. There were 5 catagories of discrimination index as stated in Table 3.4

Table 3.4 Catagories of Discrimination Index

Catagories	Discrimination Index
Very Bad	Negative
Poor	0.00 – 0.20
Satisfactory	0.21 – 0.40
Good	0.41 – 0.70
Excellent	0.71 – 1.00

(Arikunto, 2016:232)

In purpose of knowing the distractor effectiveness, the researcher did not use coefficient or index, it was counted manually. In general the effectiveness of the distractor was judged by considering whether or not at least 5% of the students taking the test choose the distractor as the answer (*Depdikbud*, 2005). In this study, the researcher just focused on whether or not the distractors worked on the items having test-wiseness.

The last investigation was checking the items reliability. In Iteman, the *r* of reliability can be checked in Iteman's scale statistic where Alpha was the coefficient of reliability. Checking reliability was essential in order to know the consistency of the test. The level of reliability can be seen in Table 3.5 as follows.

Table 3.5 Catagories of Reliability

Catagories	Reliability
Very Low	$0.00 < r \leq 0.2$
Low	$0.2 < r \leq 0.4$
Average	$0.4 < r \leq 0.6$
High	$0.6 < r \leq 0.8$
Very High	$0.8 < r \leq 1.00$

(Arikunto, 2016:89)

Knowing all information of items test-wiseness through analysing the test-wiseness subscales, the frequency of the appeared test-wiseness, and checking the

items analysis, such as item validity, the item difficulty, item discrimination, distractor effectiveness, and item reliability of each test-wiseness subscale of each school level, the researcher was going to draw a conclusion.

RESULTS

As explained previously, some test-wiseness subscales for ESL/EFL were found in the national examination items. There were 3 test-wiseness subscales. Those were 10 stem option, 10 similar option and 3 item giveaway. Thus, this is clear that both stem option and similar option are the most frequently appeared test-wiseness subscales, followed by item-giveaway.

Table 4.1 The Frequency of Test-wiseness Subscale in National Examination

No	Year	∑ items having Test-wiseness for subscale:			
		Stem option	Grammar Cue	Similar option	Item giveaway
1	2018	4	0	4	0
2	2017	5	0	3	2
3	2016	1	0	3	1
	Total	10	0	10	3
	Maximum	150	150	150	150
	Percentage	6.7%	0%	6.7%	2%

The frequency of test-wiseness in the national examination items of 2016 was 10% where there were 8 out of 50 items possessing test-wiseness. In national examination of 2017, the frequency of test-wiseness appearance was 20% or 10 of 50 items having test-wiseness. Moreover, the frequency of test-wiseness appearance in national examination of 2018 was 16% or 8 of 50 items. In conclusion, the frequency of test-wiseness in the national examination items was 15.3%.

Table 4.2 The Frequency of Test-wiseness in National Examination Items of English Subject

No	Year	∑ Items	∑ of Items Having Test-Wiseness	Percentage (%)
1	2018	50	8	16%
2	2017	50	10	20%
3	2016	50	5	10%
4	All years/Total	150	23	15.3%

Then, the findings based on item analysis showed that the items possesing test-wiseness done by all level school students had high validity, except similar option for lower level students. Generally, the items were also easy for higher and middle level school students, but the lower level found a little bit difficult to do

all items of all subscales except items of item giveaway. Besides, averagely the items possessing the test-wiseness had good ability to discriminate students' competence. Moreover, in general it seems like the distractors appear to be not plausible so that the students could get the key easily, especially for the students from higher and middle level school.

Lastly, reliabilty of the items on stem option and similar option which were applied for all level schools students were highly reliable, This means that the items possessing test-wiseness on all test-wiseness subscales gave cosistent score for all kinds of the test-takers.

Table 4.8 The Summary of Test-wiseness Statistical Analysis

No	School	Test-wiseness Subscales	Validity		Item Difficulty		Item Discrimination		Distractor Effectiveness	Reliability	
			Mean Score	Catagory	Mean Score	Catagory	Mean Score	Catagory	Percentage of items with Unworked Distractor	Score	Catagory
1	2	3	4	5	6	7	8	9	10	11	12
1	Higher Level School	Stem option	0.745	High	0.838	Easy	0.633	Good	70%	0.850	Very high
		Similar option	0.670	High	0.700	Average	0.541	Good	70%	0.722	High
		Item giveaway	0.763	High	0.875	Easy	0.492	Good	100%	0.950	Very High
2	Middle Level School	Stem option	0.916	Very high	0.809	Easy	0.750	Very good	100%	0.916	Very High
		Similar option	0.773	High	0.886	Easy	0.616	Good	100%	0.781	High
		Item giveaway	0.778	High	0.971	Easy	0.497	Good	100%	0.989	Very High
3	Lower Level School	Stem option	0.621	High	0.517	Average	0.750	Very good	40%	0.635	High
		Similar option	0.446	Sufficient	0.424	Average	0.616	Good	30%	0.149	Very Low
		Item giveaway	0.763	High	0.598	Average	0.497	Good	100%	0.812	Very High

DISCUSSION

As previously discussed, it was shown that several types of test-wiseness subscales for EFL/ESL exist in a multiple choice test, even the standardized ones. As used by Allan on the study of test-wiseness of reading test in the area of ESL/EFL,

the four major subscales were considered in this study. They were stem option, grammar cue, similar option and item giveaway.

It has been explained in results that three test-wiseness subcales do exist on the national examination items, except grammar cue. This is possibly because nowadays, the objectives of learning English is not focused on the grammatical knowledge, but developing the students' English competence, i.e. students' oral and written skills. Since they do not learn grammar specifically, but contextually, then the assessment should be in line with the learning process. It is the possible reason why grammar cue is not appeared in this study. Another subscale is item giveaway. The subscale has very small portion, whereas the other two subscales seem dominate the items possessing test-wiseness, those are stem option and similar option. Both of them are the most frequently appeared in the examination items in the three sets of examination items.

Based on the presented results, it shows that the test-wiseness appears every year, although the frequency is different. The range is between 10% to 20% for each. Totally, the items possessing test-wiseness in the three sets of national examination items are 15,3%. Thus, it is very clear that the students can find some national examination items possessing test-wiseness, eventhough such kind of items is in small portion.

The national test developer may not consider test-wiseness in creating the national examination items. As a result, it still has small portion of test-wiseness. However, the subscales appeared in the examination items seem to be reduced. This may be because *Puspendik* choose experienced teacher to write the examination items. They have good knowledge of how to create good multiple choice items. This may be the result of good continous professional development of teachers enforced by the ministry of Education, which enhances the teachers to enlarge their knowledge of paedagogical aspects, especially knowledge of creating various kinds of tests. Moreover, the small number of test-wiseness in examination items is also possibly caused by series of steps before the test items are administered in the examination, as explained in *Pedoman Penulisan Soal*. The steps are 1) providing the blueprint; 2) writing items; 3) reviewing and revising; 4) assembling the items; 5) trying out the items; 6) analysing the try out result; and 7) selecting the items (*Puspendik*; 2-3). This is in line with Salkind opinion (2013) that creating multiple choice items consume much times since it needs revision over revision to make it better.

Based on the findings that the items are easy for higher and middle level school students, but the lower level found difficult for all items of all subscales except items of item giveaway. From the explanation, it seems that there is a tendency that the students are able to recognize and take advantage of test-wiseness.

The statement of Otoum, et al (2015) should be considered that not all students know how to deal with the test situation. In other words, some of them do not know the test-wiseness strategy. Furthermore, Rahadi (2001) says that there are differences between the percentages for highest achievement students and lowest achievement students when using test-wiseness skills. In line with Rahadi, the findings of the study shows the similar thing.

Otoumn et al. (2015) explains further on the conclusion of their study that there is a collection of cognitive abilities that the individual employs in the test regardless of the knowledge content of the studying material, and it is considered as an important source of variation in the scores and also it is a factor that explains the differences between students which are in the same ability and level. The test-wiseness strategies were highly used by students, and the differences were found due to the achievement level.

The previous explanation clarifies the difference between the correct answer of students coming from higher and middle level school. Generally, students coming from higher level school achieved better than those from the middle, and the those from the middle achieve better than the lower. The data shows that correct answer of both higher and middle schools are high. It means both of them have highly link with the items having test-wiseness. Logically, the higher level school should have better result, but in fact, middle level school got better result.

The teachers of higher school level explained that the students who did the test in this study comes from an upper class, an average class and a 4-semester-class. The teacher of middle level school said that she took 2 upper classes and an average class. It means the basic ability of them are similar, good English mastery.

The conclusion of Wanatabe's study (2004) states that the students of higher grade is more proficient in English than those of lower grades. The 4-semester class is special class. They can take the study of junior high school only in 4 semesters. It means that actually the students should be at grade 8, but because of the program, in the second year they can be at grade 9. It may be one of the cause why the result of the middle level school is better than the higher.

Another reason, possibly because of intensive training on how to do items containing test-wiseness. It had been done since early March for the middle level school, while the higher level school students got any intensive training only a few weeks before the research was conducted. It is in line with Yousef's study (2004) and Scharnagl (2004) that the experimental group which is trained in doing standardized test achieved better than the control group who got no training.

For the lower level school students, it is clear that their ability in English is lower than upper and middle school level students. That is why the students got less number of correct answer than the two others. This supports conclusion of Otoumn et al (2015). In other words, some of them do not know the test-wiseness

strategy. Thus, the false answer of the students coming from lower level students because of they do not have enough ability in English as well as do not know how to treat items possessing test-wiseness.

Thus, it is clear why the majority of the items are easy for students of higher and middle school, whereas they are difficult for those coming from lower level school students. In addition, the distractors of the items do not work well for students from all level schools on all subscales except the distractor of stem option and similar option items done by the students from lower level students. Therefore, the items are able to discriminate the test-takers well.

Furthermore, the results show that the items possesing test-wiseness done by all level school students have high validity, except similar option for lower level students. This is different from Watanabe's (2004) opinion that test-wiseness is a source of invalidity of tests. The result shows that only items of item giveaway subscale have sufficient validity, not low validity. Thus, this study against the idea that items possessing test-wiseness always lead to low validity for students of all level schools. This study proves that the items having test-wiseness still have possibilty to have high, even very high validity.

The statistical analysis in this study shows that the items having test-wiseness does not have any problems with the validity except the similar option subscale for the students from lower level school. The difficulty appears only on both stem option and similar option subscales for the students of lower level students. Discrimation index has no problems, whereas the distractors are mostly ineffective except for the students of lower level students. In term of reliabilty of the items on all subscales which are applied for all level schools students, the values are highly reliable, except similar option items done by the students coming from lower level school. It shows the highly consistency result of the items except one test-wiseness subscale, similar option, for the lower ability students.

In conclusion, this study goes along with opinion stated by Otoum et al. (2015) that the appearance of test-wiseness affects the students of high level school who have good cognitive ability, while the students of the lower level school seem not to have enough background knowledge to deal with the test-wiseness.

CONCLUSION

This study analyses the test-wiseness of three sets of national examination items for junior high school students, the year of 2016, 2017 and 2018. Based on the analysis, the researcher draws several conclusions.

Firstly, this research finds there are three test-wiseness subscales for EFL/ESL students, out of four, in the three sets of national examination items. Those are stem option, similar option and item giveaway.

Secondly, this research finds that test-wiseness appears in the three sets of national examination items in various portion for each year, between 10% up to 20%. Total test-wiseness of the three sets of the examination items are 15,3%. It is considered as a small portion. Thus, the frequency of test-wiseness in national examination items in the three years are small.

Lastly, the statistical analysis in this study shows that the items having test-wiseness does not any problems with the validity except the similar option subscale for the students from lower level school. The difficulty appears only on both stem option and similar option subscales for the students of lower level students. Discrimation index has no problems, whereas the distractors are mostly ineffective except for the students of lower level students. However, the reliability is high for all level school students, except the reliability of stem option of lower level students. Thus, based on the statistical analysis, it infers that the appearance of test-wiseness affects the students of high level school who have good cognitive ability, while for the students of the lower level school, it seems they do not have enough background knowledge to deal with the test-wiseness.

Moreover, this study proves that test-wiseness do exists in the national examination items, although it is in a small frequency. Therefore, for better quality of national examination in the future, it is needed to thouroughly examine the existence of test-wiseness on the items.

Based on this analysis, the researcher offers some suggestions as follows. For the test developer of national examination and all English teachers, the researcher suggests that it is very essential to obey the procedure of writing test items and recheck items in order to avoid test-wiseness appearance, especially the three test-wiseness subscales: stem option, similar option, and item giveaway.

In addition, it is also suggested for the headmaster of junior high school to hold a teacher training on developing the teachers'skill to construct multiple choice test, especially to avoid test-wiseness, and to hold an extra lesson for the ninth graders in purpose of coaching the students to prepare national examinantion for some of the items are still containing test-wiseness.

Furthermore, for the future researcher, analysing the test-wiseness of the future national examination items with broader population is still worthly needed to make sure the good quality of national examination items. In addition, the next researcher may investigate how to deal with the test situation for the test-takers.

REFERENCES

Allan, A. 1992. Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing*, (Online), Volume 9 Number 2, (http://ltj.sagepub.com/content/9/2/101 DOI: 10.1177/026553229200900201, retrieved on 12 October 2018)

Arikunto, S. 2016. *Dasar-dasar Evaluasi Pendidikan Edisi Kedua*. Jakarta: Bumi Aksara.

Ariyana, L. T. 2011. *Analisis Butir Soal Ulangan Akhir Semester Gasal IPA Kelas IX SMP di Kabupaten Grobogan.* Unpublished Thesis. Semarang: Mathematic and Science Faculty, State University of Malang

Badan Standar Nasional Pendidikan. 2018. *Prosedur Operasional Standar (POS) Penyelenggaraan Ujian Nasional Tahun Pelajaran 2018/2019.* Jakarta: BSNP

Brown, H.D. & Lee, H. 2015. *Teaching by Principles An Interactive Approach to Language Pedagogy (Fourth Edition).* New York: Pearson Education Inc.

Depdikbud. 2005. *Panduan Analisis Butir Soal Pilihan Ganda.* Jakarta: Balitbang-Depdikbud

Evans, W . 1984. Test Wiseness: An Examination of Cue Using Strategies. *The Journal of Experimental Education*, (Online), Volume 52, Number 3, (http://dx.doi.org/10.1080/00220973.1984.11011883, retrieved on 20 October 2018)

Millman, J., Bishop, C.H., & Ebel, R.1965. An Analysis of Test-Wiseness. *Educational and Psychological Measurement*, (Online), Volume XXV, number 3, (http://epm.sagepub.com/content/25/3/707 DOI: 10.1177/001316446502500304, retrieved on 20 October 2018)

Otoum, A, Khalaf,H.B., Bajbeer, A.,Hamad, H.B. 2015. The Level of Test-wiseness for the Students of Arts and Science Faculty at Sharourah and Its Relationship with Some Variables. *Journal of Education and Practice*. 6(29): 102-113

Peraturan Menteri Pendidikan dan Kebudayaan R.I no 23/2016 tentang Standar Penilaian Pendidikan. 2016. Jakarta: Kementerian Pendidikan dan Kebudayaan

Powell, R.R. 2006. Evaluation Research: An Overview. *Library Trends,* 55(1): 102-120.

Pusat Penilaian Pendidikan. 2017. *Panduan Penulian Soal 2017 SMP/MTs.* Jakarta: Kementerian Pendidikan dan Kebudayaan

Qimala. 2013. *Test-Wiseness Item Analysis On SMAN 1 Kediri's Selection Tests In The English Subject.* Unpublished Thesis. Malang: Faculty of Letter, State University of Malang.

Rogers, W. & D. Bateson. 1991. Verification of a model of test taking behavior of high school seniors. *Journal of Experimental Education,* 59: 331-349

Sulistyo, G.H. 2002. *Language Testing Some Selected Terminologies and Their Underlying Basic Concepts.* Unpublished Textbook. Malang: English Department, State University of Malang

Sulistyo, G.H. & Rahmajanti, S. 2003. *Tes Bahasa Inggris Sekolah Dasar.* Malang: Bayumedia

Sulistyo, G.H. 2018. *EFL Learning Assessment At Schools An Introduction to Its Basic Concepts and Principles*. Malang: CV Bintang Sejahtera Abadi

Seaman, P. 2003. *Multiple Choice Testing: will it work for Me?*, (Online), (http://www.lib.unb.ca/Texts/Teaching/JAN03/seaman.html, retrieved on 14 February 2019).

Scharnagl, T.M. 2004. *The Effects of Test-taking Strategies on Students' Reading Achievement.* Unpublished Doctoral Dissertation. Union Institude and University, UMI 31444027

Watanabe, Y. 2004. Exploring Test-Wiseness of Japanese Senior High School Students. *Japanese Journal.* 59: 27-34.

Yousef, Imad. 2004. The Effect of The Test-wiseness on Collection of a Sample from Students of College of Education-Minia University. *Education and Psychology Research Journal.* 17(3): 349-383